ARL-TR-7705 ● JUNE 2016

US Army Research Laboratory

Complex Event Processing for Content-Based Text, Image, and Video Retrieval

by Anne-Claire Boury-Brisset, Elizabeth K Bowman, Gertjan Burghouts, Barbara D Broome, John Duselis, Bruce Forrester, V Melissa Holland, Jonathan Howe, Jasper van Huis, Peter Kwantes, Bhopinder K Madahar, Adem Yaşar Mülayim, Raghuveer M Rao, and Douglas Summers-Stay

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-7705 ● JUNE 2016

US Army Research Laboratory

Complex Event Processing for Content-Based Text, Image, and Video Retrieval

by Elizabeth K Bowman, Barbara D Broome, V Melissa Holland, and Douglas Summers-Stay
Computational and Information Sciences Directorate, ARL

Raghuveer M Rao
Sensors and Electron Devices Directorate, ARL

John Duselis
US Air Force Research Laboratory, Dayton, OH

Jonathan Howe and Bhopinder K Madahar
UK Defence Science and Technology Laboratory, Porton, Salisbury, UK

Anne-Claire Boury-Brisset and Bruce Forrester
Defence Research and Development Canada, Valcartier, Quebec

Peter Kwantes
Defence Research and Development Canada, Toronto, Ontario

Gertjan Burghouts and Jasper van Huis
TNO, The Hague, Netherlands

Adem Yaşar Mülayim
Atos Turkey, Ankara, Turkey

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
June 2016	Final	October 2014–September 2015

4. TITLE AND SUBTITLE
Complex Event Processing for Content-Based Text, Image, and Video Retrieval

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Anne-Claire Boury-Brisset, Elizabeth K Bowman, Gertjan Burghouts, Barbara D Broome, John Duselis, Bruce Forrester, V Melissa Holland, Jonathan Howe, Jasper van Huis, Peter Kwantes, Bhopinder K Madahar, Adem Yaşar Mülayim, Raghuveer M Rao, and Douglas Summers-Stay

5d. PROJECT NUMBER
ET 086

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-CII-T
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-7705

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
NATO-CSO
7 rue Ancelle
92200 Neuilly sur Seine
France

10. SPONSOR/MONITOR'S ACRONYM(S)
NATO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report summarizes the findings of an exploratory team of the North Atlantic Treaty Organization (NATO) Information Systems Technology panel into Content-Based Analytics (CBA). The team carried out a technical review into the current status of theoretical and practical developments of methods, tools, and techniques supporting joint exploitation of multimedia data sources. In particular, content-based information retrieval and analytics was considered as a means to allow military experts to exploit multiple data sources in a rapid fashion for sensemaking and knowledge generation. Elements included contextual understanding of complex events through computational/human processing techniques, event prediction through the automated extraction of network features, temporal trends, hidden clusters and resource flows, and the use of machine processing for automated translation, parsing, information extraction, and summarization of unstructured and semistructured data. The main conclusions of the study are that important research gaps exist in all the technical areas covered in this report. Though the research areas and developments are being advanced in the military sector and the civil sector, in particular, they remain at low levels of technical maturity for defense and security system applications. It is recommended that this NATO collaborative research effort be expanded to advance those approaches that are most pertinent to our overall aim of enhancing the contextual understanding of complex events through CBA of heterogeneous multimedia streams.

15. SUBJECT TERMS
content-based analytics, multimedia data sources, sensemaking, knowledge generation, event prediction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Elizabeth K Bowman
Unclassified	Unclassified	Unclassified	UU	42	19b. TELEPHONE NUMBER (Include area code) 410-278-5924

Contents

List of Tables

Acknowledgments

The authors are grateful to all members of the exploratory team and for the support provided by the defense and security research programs in each of their organizations, including the provision of relevant program/project technical information. In particular the authors would like to express their thanks to Dr Peter Harvey, Defence Science and Technology Laboratory (DSTL), for his contributions to the meetings and technical discussions. They would also like to thank Colonel (Rtd) Philippe Soète, Executive Officer to the Director and Interim Information Systems Technology Panel Executive, Collaboration Support Office, North Atlantic Treaty Organization Science and Technology Organization for superb organizational support and technical report guidance, and Ms Brittany Mckinnon, US Army Research Laboratory (ARL), for technical report editing and administrative support. Finally, they would like to acknowledge the leadership and technical guidance provided by the leaders of the exploratory team, Dr Barbara D Broome, ARL, and Professor Bhopinder K Madahar, UK DSTL.

1. Introduction

1.1 Context

In today's coalition military environment, such as North Atlantic Treaty Organization (NATO) capabilities and operations, decisions are needed quickly within a contextual environment that is increasingly uncertain and complex. Current military information systems do not collect sufficient data on local attitudes, culture, and human issues, all of which frame the military decision parameters. Analysis of information tends to be centralized in operational or strategic levels of command, leaving tactical commanders without the ability to maintain awareness of information that is localized and dynamic.

Content-based information retrieval and analytics, referred to as Content-Based Analytics (CBA) in this report, is a means to allow military experts to exploit multiple data sources in a rapid fashion for sensemaking and knowledge generation. Elements of these processes include contextual understanding of complex events through computational/human processing techniques, event prediction through the automated extraction of network features, temporal trends, hidden clusters and resource flows, and the use of machine processing for automated translation, parsing, information extraction, and summarization of unstructured and semistructured data.

This report is the result of scientific exchanges that occurred under the NATO Information Systems Technology (IST) panel to provide an overview of the current state of the art in the field of CBA. We considered research that creates, harvests, and exploits technologies that can help realize the potential for improved decision making without imposing a need for increased Warfighter numbers or their cognitive workload.

1.2 Objectives

Within this context, the main objective was to review, through collaboration among NATO partner scientists, the development of theoretical and algorithmic tools supporting joint exploitation of multimedia data sources. We considered 3 multimedia types in the activity period to include image, video, and text, with voice as an additional possibility depending on the progress achieved with the first 3. Activities included 1) defining scope and extent of investigations, 2) developing an approach that includes identifying the extent to which a common framework can be developed for representing data from the different classes, 3) reviewing

machine/deep learning (DL) tools to automatically detect and classify events from combinations of one or more data classes, and 4) data collection and ground truth labeling.

Investigation into developing methods for joint exploitation of multiple media requires effort at several different levels. These range from identifying real-world scenarios where 2 or more multimedia types co-occur to designing features that best lead to efficient extraction of actionable information. Research in this area is still in its infancy. Given that military applications, especially those that are intelligence related, present challenges of varying degree and are more involved than applications in the civilian world, it is accordingly necessary that the military scientific community undertakes extensive effort in developing solutions to research problems in the field. It also stands to reason that collaborative research among NATO partner scientists can contribute significantly to the development of these solutions. However, the important technical enablers need to be known and are discussed in this report.

Some technical enablers are known by the NATO partner scientists from previous and ongoing research. In the text analytics (TA) domain, the review effort concentrates on resolving references to multiple objects that share a common name or extracting references to events or relationships between objects. Current work on techniques for improving the automated processing of text by restricting the domain show some promise for improving both the accuracy and timeliness of automated tools for natural language processing (NLP). Further, they provide a foundation for longer-term goals of automating causal analysis, inference of human intent, and extraction of implicit dynamics from text.

Text and video analysis for extracting information to support content cannot be done in isolation. NATO and coalition military leaders, commanders, and intelligence analysts need interoperable tools that cross-cue knowledge obtained from one method to generate taskings in the other. Hence, a focus on building the cross-cued solution from advances in video and TA is required.

1.3 Structure

The results of the technical review undertaken by the exploratory team (ET) are summarized in this report. The following section provides an overview of the state of the art in a number of technical areas considered to be of relevance and emerging (i.e., low levels of technical maturity compared to others in the report). The section also discusses the progress on developments of common frameworks for multimedia analytics and the availability of data sets for development and testing.

The report is completed with sections on research gaps, promising methods, and conclusions and recommendations as outlined by the findings of the team.

2. State of the Art

2.1 Text Analytics

2.1.1 Background

TA is loosely defined as the ability to distill information from textual characters. It is a multidisciplinary research area spanning, but not limited to, linguistics, psychology, engineering, physics, mathematics, and computer science. Tokenization, or segmentation, is the process of grouping the characters into units (words). After a text is tokenized and tagged, and drawing on inherent world knowledge, words are grouped into phrases, sentences, paragraphs, and documents to derive meaning. The next step is to apply statistical properties to start classifying the different groupings of words. Methods include probabilistic modeling and statistical inference using techniques such as Hidden Markov Models, Mixture Models, Bags of Words, Artificial Neural Networks (ANNs), higher-order logic, graph theory, and classifiers. Most are well known in the broad research areas of machine learning and pattern recognition (Duda et al. 2000).

Further inspection of the words, sentences, paragraphs, and documents reveal more information and can start deriving meanings, intent, and higher-level information abstractions. This is where ontologies, lexicons, and taxonomies come into consideration. These data help structure the interpretation of the text to derive meaning. The complexity of analyses rises dramatically as does its uncertainty when attempting high abstraction (Bird et al. 2009)

A key to TA success is the use of corpora for training and model building. A corpus consists of data that has some truth behind it and is relevant. Many corpora are used to allow the ontological mapping of a language to derive meaning. They provide the ability to build models of grammar and syntax for word and sentence structure. Size of the corpora must be significant for the models and statistical methods to base decisions when analyzing the data for high-level meaning. However, the model is only as good as its training data. Therefore, text analyses in an increasingly important language(s) may suffer until more data are compiled in it.

As layers of abstraction are built, gaining meaning from text is called information extraction (Bird et al. 2009). This level can have multiple applications such as semantic analysis, understanding, intelligence gathering, and of recent great interest, media analysis. It relies heavily on machine-learning techniques to discover interesting aspects of the text (unsupervised) or increase understanding of text in focused investigations (supervised). More on supervised and unsupervised learning can be found in Duda et al. (2000). Information extraction (IE) includes techniques such as entity extraction, relationship recognition and building (between words and parts of speech), and higher-order representation of text (such as trees versus tokens). Trees are of great importance because the hierarchical structural representation that results from analysis of the text forms the basis for interpreting the meaning of the text, such as media analysis (Grimmer and Stewart 2013).

2.1.2 Examples of Existing Text Analytics Programs

Several US Defense Advanced Research Projects Agency (DARPA) projects have addressed features of TA. The Low Resource Languages for Emergent Incidents focused on dramatic advancement of the state of computational linguistics and human language technology to enable rapid, low-cost development of capabilities for low-resource languages (e.g., those languages for which no automated human language technology capabilities exist). The Deep Exploration and Filtering of Text (DEFT) program seeks to address information implicitly, but not explicitly, shared in exchanges. Automated, deep NLP technology may provide the solution for more efficient processing of text information and enabling understanding connections in text that may not be readily apparent to humans. The DEFT project also aims to enable a capability to integrate individual facts into large domain models to support assessment, planning, and prediction. The Broad Operational Language Translation program is aimed at enabling communication with non-English-speaking populations and identifying important information in foreign-language sources. Specific functions include 1) allowing English-speakers to understand foreign-language sources of all genres, including chat, messaging, and informal conversation; 2) providing English-speakers the ability to quickly identify targeted information in foreign-language sources using natural-language queries; and 3) enabling multiturn communication in text and speech with non-English speakers. The Multilingual Automatic Document Classification, Analysis and Translation (MADCAT) program is to automatically convert foreign language text images into English transcripts, thus eliminating the need for linguists and analysts while automatically providing relevant, distilled actionable information to military command and personnel in a timely fashion. MADCAT technologies are able to analyze images to determine language and type of script; classify images to determine the kind of material that each image presents (photo, newspaper article,

technical memo, ledger, etc.); and segment images and interpret different text zones, including classification and parsing of tables, produce transcripts of images in their source languages, whether printed or handwritten, and produce accurate English translations of source language text. The program has developed optical character recognition and machine translation capabilities for 11 languages: Arabic, Chinese, Dari, Farsi, Hindi, Pashto, Spanish, Russian, Thai, Urdu, and Korean.

The US Intelligence Advanced Research Projects Activity (IARPA) invests in high-risk, high-payoff research programs to tackle some of the most difficult challenges in the intelligence community. Several IARPA programs have advanced the state of the art in TA in recent years. The Metaphor program was an attempt to exploit the use of metaphors by different societies to gain insight into their cultural norms. Shared concepts and worldviews of a society's members are difficult because these tend to be implicitly shared and are often hidden from an outsider's view. Interpreting metaphors can help decision makers to be effective in understanding how beliefs influence a person's approach to a complex topic and how they might behave in a situation. Research into metaphors has uncovered inferred meanings and worldviews of particular groups or individuals. These include the characterization of disparities in social issues and contrasting political goals, exposure of inclusion and exclusion of social and political groups, and understanding of psychological problems and conflicts. The IARPA Finder program was developed to apply geo-location tags to image/video scenes of interest. The program seeks to develop innovations that include the 1) integration of analysts' abilities and automated geo-location technologies to solve geo-location problems, 2) fusion of diverse, publicly available, but often imperfect data sources, and 3) expansion of automated geo-location technologies to work efficiently and accurately over all terrain and large search areas.

2.2 Video Analytics

In the area of video analysis, the focus of automation technologies has been on providing a foundation for activity identification. Our ability to retrieve video segments or chips from large video streams based on what is inside the clip is extremely limited. Video sources rarely provide more than a minimal amount of metadata (day, time, location, source) that will allow the user to filter out streams of irrelevant information. While we have had some success in identifying objects within a video stream, automatically identifying and marking-up activities would greatly enhance the decision maker's ability to take full advantage of the real strength of video, its recording of change over time. Furthermore, such developments could enable high performance "frame-rate" computations and faster analysis of ever-increasing volumes of streamed video data.

2.2.1 Detection, Extraction, and Tracking of Objects

There is a military need for automatic video analysis to select just the relevant parts within the increasingly large amounts of collected video data. Scarce human analyst resources can then be used for analyzing the only the selected parts, such as human activity that may pose a military threat. This requires algorithms that are able to analyze video data (e.g., to detect and track people and vehicles and to identify where a particular human activity is taking place).

In the US DARPA Mind's Eye program, the aim was to develop a robotic scout, endowed with a camera that could recognize 48 activities and generate reports about what it had detected. The activities of interest varied from running to picking up an item to 2 persons exchanging an item. The actors were persons and vehicles. The program finished in 2013 and as a result, it produced new algorithms. The novelty mainly was that activities were recognized by identifying the agent carrying out the activity and its interactions with other people and objects in the scene. In the Netherlands at an airport setting, the technology was further advanced to consider subtle activities involved in complex events such as theft. Such events are hard to recognize because they are only partially visible and involve interplay of multiple agents. A prototype is running live on one camera. The next step for such algorithms is to extend them to multicamera analysis, in order to associate observed parts of an incident across different cameras. This project, named HARVEST (Human Activity Recognition in VidEo STreams), is based on a supervised learning system where behavior models are created from manual annotations of events in video streams. The DARPA program algorithms were tailored to the airport environment and selected incidents. The prototype has been in place since July 2015, after 8 months of development.

Often, military video data are collected by moving and aerial platforms. This adds the challenge of analyzing the scene and actors while the camera is moving and sometimes very distant from the scene. Many algorithms have been developed to preprocess such videos to stabilize and enhance the image quality. First, research steps have been taken to analyze the contents automatically, such as detecting and tracking people and vehicles. Tracking is hard because the person may be visible for a short time only, the zoom may vary significantly, and the source (color and infrared) may be toggled. Automated analysis of activities is challenging for the same reasons. For short-term and simple activities of 1 or 2 persons (e.g., run and dig), promising results have been achieved. The next steps are to improve the accuracy of the algorithms and to extend identification to more complex activities. During the last 10 years, wide area motion imagery (WAMI) sensors have become available, which enable the collection of video data over multiple square kilometers

from an aerial platform. Due to the large quantities of data that WAMI generates, it is highly desirable that the contents can be analyzed automatically. Although the first results look promising, more automated information extraction also means more false alarms.

2.2.2 Patterns of Life Identification in Full-Motion Video (FMV)

Interpretation of video scenes is accomplished by analysis and classification at various levels, from the level of region and object segmentation to identification of events and actions. The term pattern of life (POL) in the FMV context refers to determining events and event sequences that are routine in its scene content. For example, FMV shots taken daily of the front area of a building may show the raising of a flag every morning at 0600. This event can then be regarded as part of the daily POL in that locale. Establishing POLs is an important step for enabling the distinguishing of normal behavior or events from those that are unusual or anomalous.

The degree of challenge in automatic determination of POLs varies with the types of events involved and the FMV setting itself. Some examples include the following:

- With stationary video cameras focused fixedly on a field where people walk, it is not difficult to determine paths normally taken by people (the POL for walking). On the other hand, where specific objects and their positions in a scene have to be identified, the challenge is significant. For example, requiring image processing and machine-learning algorithms to recognize repeated appearance of the same individual but with changing attire as normal and part of the POL is difficult.
- Platform perspective and motion, as with unmanned autonomous system (UAS)-gathered image data, is another challenge. Scaling and spatial and temporal transformations between scenes, optical flows, and occlusions are some of the technical issues that need to be addressed to reduce uncertainties in the POL.
- Time scale presents yet another type of challenge. For example, from one day to the next, a particular scene may not show much variation and yet changes may accumulate over long periods of time resulting in a shift in the POL that may or may not be anomalous. This leads to the notion of POLs as being different as a function of the time scale involved.

Hence, determining anomalies amounts to determining deviations from the POL which itself is uncertain. The US Army Research Laboratory (ARL) has been investigating unsupervised learning methods for long-term POL determination with FMV data collected as part of a DARPA Computer Science Study Group–funded effort.

2.2.3 Examples of Existing Programs in Video Analytics

- DARPA Mind's Eye (2010–2013): detection and localization of 48 human activities in a video stream (TNO Netherlands [Netherlands Organisation for Applied Scientific Research] was one of the teams)
- DARPA Video Image Retrieval and Analysis Tool (VIRAT) (2008–2012): allows users to query real-time video to create alerts for dangers
- DARPA Visual Media Reasoning (2012–2015): generating metadata about images (people, weapons, etc.) with the goal of rapidly exploiting confiscated media images to generate intelligence for counterinsurgency and counterterrorist operations
- IARPA Aladdin (2011–2014): recognition of video and audio events in very large video sets to create support for analytic needs
- NL MOD V1340 (2013–2016): experimentation with novel data analysis techniques for UASs through Concept Development and Experimentation
- NL MOD V1508 (2015–2018): generating metadata about aerial videos (FMV) by combining video analytics with context (including text/chat, geographic information systems)
- UK Ministry of Defence (MOD) Centre for Defence Enterprise (2014): themed competition: Information Processing and Sensemaking
- UK MOD Research Programmes (Current) (Intelligence and Countering Adversary Networks, Assured Information Infrastructure, Decision Support and Experimentation): portfolio of current research projects related to systems and analytics (e.g., text, images/video) and experimentation for decision support and actionable information

2.3 Concepts for Integrated Text and Video Analytics

2.3.1 Background

We would like to use natural language to retrieve, summarize, and ask questions about visual data. While this has been a goal of computer vision since its inception, early approaches used hand-designed features and knowledge bases that could not be generalized outside a very limited domain. The survey paper "Knowledge-Based Image Understanding Systems: A Survey" explores the capabilities of many of these systems (Crevier and Lepage 1997). However, the goals of these systems were mainly to make use of top-down information about a scene to make up for the low computational power and limited library of image processing tools available at the time. In "Automating Knowledge Acquisition for Aerial Image Interpretation", for example, hand-built detectors for roadways and buildings in aerial images of airports made use of knowledge about the connectivity between terminal buildings and roads to recognize the function of particular objects in airport images (McKeown et al. 1989). Efforts on integrating world knowledge with image processing have concentrated more on the image search problem. Instead of attempting to understand one particular image in a deep way, these methods return many images containing examples of the object searched for, and use world knowledge to understand the natural language search query. For a survey of these efforts, see "Semantics Extraction from Images" (Pratikakis et al. 2011).

In the last 5 years, machine-learning approaches have made these problems much more tractable. DL, convolutional networks, recurrent and long short-term memory networks have made it possible to recognize thousands of classes of objects in images and to generate captions for the images. For example, "I2t: Image Parsing to Text Description" uses an intermediate web ontology language (OWL) representation to go from object recognition to sentence generation (Yao et al. 2010). "Show and Tell: A Neural Image Caption Generator" (Vinyals et al. 2015) and "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)" (Mao et al. 2015) also use neural networks to generate captions describing images with some success at creating captions that are similar to those a human would produce. A more advanced approach "Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework" learns a model for the visual space and the linguistic space simultaneously in a joint neural model that incorporates features of both (Xu et al. 2015).

However, all of these approaches have an opaque internal representation that cannot be used to answer questions about images or modify the description based on a subject of interest. Such tasks require a more in-depth understanding of what attributes of an image are necessary for it to depict something. The Center for Language and Speech Processing at Johns Hopkins University 2012 summer workshop outlined an approach to understanding images by recognizing parts of objects, materials, and scene context (Blaschko et al. 2012). A Visual Question Answering (VQA) approach using the Microsoft Common Objects in Context (Lin et al. 2015) data set poses a problem that seems likely to be solved only by incorporating such information about the objects in the image (Antol et al. 2015). "Extracting Visual Patterns from Deep Learning Representations" introduces an approach to extracting semantic vector representations (similar to those used in linguistic representations of meaning) directly from images (Garcia-Gasulla et al. 2015). Action and activity recognition from video is an important piece of this puzzle, but results in this area are still very specialized and not applicable outside of limited domains (Girju 2003). The DARPA Mind's Eye program, for example, originally hoped to recognize dozens of verbs but in the end was able to build satisfactory detectors for only a few actions like walk, run, and pick up. However, at the Computer Vision and Pattern Recognition conference in 2015 several speakers mentioned that next year they hoped to apply these advanced neural network approaches to captioning video.

2.3.2 Examples of Existing Integrated Text and Video Analytics Programs

One of the applied research areas for integrated text and video analytics is human-robot interaction. Several programs are supported in this area, originating from foundational research of the Situation Understanding Bot through Language and Environment Multi University Research Initiative (SUBTLE MURI). Software developed under the SUBTLE MURI was transitioned to ARL where existing robotics assets were used to explore the problems of giving commands and asking questions of robotic teammates about the environment in which they are working.

Together with Devi Parikh and Dhruv Batra of Virginia Tech and 2 of their PhD students, ARL researchers are developing methods of using machine learning to answer questions about images. These include both synthetic images (about which we have exact knowledge) and a large corpus of natural images (the Microsoft Common Objects in Context data set). Given an input image and a free-form natural language question about the image, the task of the system is to provide an accurate natural language answer. The questions are open ended, as are the answers. A system that succeeds at visual question answering is more likely to have truly and

fully understood the image. The question drives the answer to be specific, making the task better posed than generic image captioning.

ARL researchers have also been working closely with 2 groups at Carnegie Mellon University as part of the ARL Robotics Collaborative Technology Alliance. One group of students developed software that can take a map and spatial prepositional phrase and return objects that fulfill the conditions in the map. For example, requesting "all people behind the building and to the left of the car" returns the identifier for all the people in the map that match that description. The software learns from a few examples for each preposition. A second group shared code to take images and 3-dimensional point clouds and identify which class everything visible belongs to buildings, grass, sky, trees, pavement, and so forth. This has been used by ARL robots for navigating safely in unknown terrain by sticking to regions we know will be safely traversable, such as roads and sidewalks. We have also been working with a PhD student from Bruce Draper's group at Colorado State to develop ways of quickly and accurately labeling training data for such classifiers.

From an international perspective, NATO commissioned a study to explore how social media might be exploited for military purposes. While social media initially was primarily text, video and imagery is gaining popularity as a sharing mechanism. These information sources provide a suitable foundation for the study of text and video analysis for military understanding of complex environments. In recent years, the culmination of mobile networking technology, ubiquitous network availability, and the public's willingness to openly share information in open sources has delivered an unprecedented availability of data. To identify how social media might be useful for military purposes, NATO commissioned the System Analysis and Studies (SAS)/IST Research Technology Group (RTG) 102. SAS/IST RTG 102 was organized in November 2013 to identify rational and effective ways of exploiting social media as a source for intelligence purposes. A further goal was to investigate potentially relevant methods and tools for use by intelligence analysts when exploiting social media sources. The group is composed of Open Source Intelligence (OSINT) practitioners and defense scientists, each having specific roles in achieving the following goals.

For OSINT practitioners, the need from social media was identified as a tool/script, or methodology for developing a real-world scenario, or Request for Information, that is relevant/timely for their own country's priorities. They were responsible for determining the capabilities of various open source tools and identifying additional capabilities that might be required for military operations. Their counterparts, the defense scientist members, were responsible for performing research and validation on selected tool/script, or methodology, using the US-provided ground-truth data set to determine how well the social media platform met OSINT needs,

and exploring future capabilities for intelligence requirements. Together, the practitioners and scientists worked to define strategies to access, monitor, analyze, and estimate future states from a variety of social media data. Accessing social media data requires initial capture and storage of data, to include applying anonymity measures to data. Monitoring the range of social media platforms is a challenge due to the dynamic range of platforms and the rapidity with which common platforms modify their procedures for accessing data. Analysis of social media data is a rich endeavor as platforms increasingly expand their capabilities, including geo-location of messages, sentiment, social networks, and temporal aspects of information sharing. For social media to be a valuable intelligence capability, estimation will need to be further developed, with ways in which to detect deception and social bias.

As of summer 2015, efforts by the group have concentrated on text analysis of social media concentrating on Twitter as a source. Little work has been done on video exploitation, not because it is not pertinent and required, but because it is much harder than text analysis. Even though, given the nature of social media, text analysis is very difficult in its own right.

There are millions of hours of video online, and this, in and of itself, makes video exploitation difficult. An initial approach would be to develop tools and methods that would enable the filtering and finding of relevant videos. One way to accomplish this would be using the existing TA to parse through the comments and descriptions that are part of the metadata typically found attached to online video. The NIST TRECVID is a useful effort to refer to on a continuing basis (http://trecvid.nist.gov/). As the SAS/IST RTG 102 moves forward, the group will maintain close ties with the IST 086 effort to determine points of leverage between the groups.

2.3.3 Human-Machine Collaboration

The goal of automating text and video extraction for content understanding will not remove human operators from the decision-making loop. Human interaction with and understanding of the text and video analytics process is paramount to developing trust in results and viewing the computational process as a valued addition to the sensemaking endeavor. In today's military environment, teams are increasingly operating across wide geographic areas, creating a need to understand how computational elements can improve team collaboration and sensemaking in complex settings.

Military Warfighters and intelligence analysts have to deal with a plethora of information from various sources in their operational environments. It is known that most of collected and archived data are never accessed due to poor/limited search tools, and they are not analyzed to derive insights due to limited analysis support tools. On the other hand, the new generation of military operators is familiar with mobile devices and applications that are user-friendly, and they are used to multimedia processing (e.g., take a picture/video, tag it, and publish it).

In this context, future automation in support of intelligence and operations should exploit all available information sources and consider these human characteristics for the provision of tools to provide the following:

- Simple and intuitive ways to express information requirements (e.g., content-based, faceted-search, or example-based)
- Relevant information to users that best meets these requirements from any source (e.g., observation report, image, video clips, social media) based on content and not only on limited metadata
- Information to users as soon as possible (through alerting/notification, mode push vs. pull) while not overloading the user with irrelevant information

For this to occur, a rich user context has to be defined (user role, mission objectives, area of interest, domain of interest, etc.) and maintained so that human-computer interaction is adapted to users' profiles and requirements.

2.3.4 Collaborative Sensemaking

As mentioned in Section 2.3.3, teams of human analysts are increasingly operating in geographically dispersed regions and demand supportive computational tools. In this domain, the idea of collaborative sensemaking (CS) refers generally to the processes by which members of a team work together to generate and maintain a shared understanding of a situation, problem, or solution in an effort to reach a common goal. As an area of scientific pursuit, CS is fairly nascent—presumably owing to the fact that many of the technological enablers for CS are relatively new. Umpathy (2010) outlined what he considered to be the broad but critical requirements to support CS that can be distilled down to a few basic guidelines to steer ongoing scientific work on CS processes and technologies. Generally speaking, effective CS requires the following:

- A means by which knowledge can be explicitly represented. For example, knowledge should be externalized through means such as a visual display.
- An environment, either physical or virtual, where a shared representation of knowledge can be constructed, shared, and maintained.
- A means by which team members can communicate agreement and conflicting views with respect to others' understanding or treatment of information.
- A moderator to maintain focus and effective interaction.
- A means by which information (e.g., documents) can be exchanged, categorized, and annotated (Paul and Morris 2011).

For the most part, technologies to support CS satisfy the above guidelines by providing users with a common set of tools: messaging to support communication (especially if users are distributed), a repository where relevant documents are stored, and most of the time, some kind of display tool that provides users with a shared visual representation of information relevant to their tasks.

When we consider the more specific properties of the available tools, and how they differ, it is quickly apparent that developers and researchers use the term "collaborative sense-making" differently. In what follows, we outline some of the different dimensions across which technologies differ. The classification is simple and qualitative, but for the most part, most of the tools we have discovered through our searches can be characterized by a profile across the following dimensions.

- *Visualization vs. No Visualization:* Despite the ubiquity with which CS tools use visualization as a component capability, not all collaborative enablers rely on it. For some, CS may constitute nothing more than creating a common document either asynchronously, as with writing or contributing to a wiki, or synchronously using tools such as Google Docs.
- *Collocated vs. Distributed Collaboration:* Some tools are clearly meant for collocated analysts because the systems are designed with a single interface, like a touch table, around which a team would gather (Wallace et al. 2013). Other tools, however, assume that team members are distributed either in time (e.g., Heer and Argrawala 2008; Fisher et al. 2012) or space (Goyal et al. 2013).

- *Shared or Independent Interfaces:* The extreme example of a tool for collocated collaboration is also an example of a tool for which every user interacts with the same interface (Grant 2001). The alternative, is to have users working with interfaces (either from the same or different software applications) on their own device (PC, tablet, touch table).
- *Shared Displays Are Generated by the Analytic Tools vs. Shared Displays Are Generated by Collaborators:* In most cases, CS tools possess processes for analysis as well as ones that generate one or more shared displays. This is especially true for collaborative technologies that rely on interactive visual displays of information to support analysis (i.e., visual analytic tools). The alternative is one in which analysts might use different tools for their portion of the sensemaking activity, and that the integration of knowledge across team members is done at the shared display layer. For example, CMap tools (http://cmap.ihmc.us) allows users to collaboratively build a concept map that serves as a shared visual representation of a knowledge model.
- *Applications with Artificial Intelligence (AI) Support to Sensemaking vs. Applications Without:* As examples of AI-supported CS tools, Co-OPR (Tate et al. 2006) is a DARPA-funded sensemaking system supported by ontological reasoning. See also Keel's EWall concept for CS that includes agents that monitor analysts' activity and infer relationships among information items being worked by the team (Keel 2007).

3. Research Gaps

It is evident from the details outlined in the previous sections that gaps in research exist in all the technical categories covered in this report. It is important to identify those gaps that are most pertinent to our overall aim. This is to enable enhancements in 1) the contextual understanding of complex events through advances in computational/human processing techniques, 2) event prediction by the automated extraction of spatio-temporal features, hidden clusters, network structures and resource flows, and 3) the application of better machine-learning/DL algorithms and processing for the automated translation, parsing, information extraction, and summarization of unstructured and semistructured data from multiple streams. If we can by example progress this through addressing the gaps in text and video analytics, then that would be not only a positive research contribution but also, because of manageable scope, realizable in systems for experimentation and testing so that value/utility can be measured using representative data sets. In parallel we would investigate the other research areas and gaps but with the aim of deepening

our technical understanding and maintaining technical currency in the latest developments.

Modeling perceptions of professional intelligence exploitation personnel is another area to be further investigated. Professional exploitation personnel can detect important events, for example, more correctly and faster than anyone else can. In doing this successful task, they probably use their previously developed complex perception background, correlating important details of the scene, knowing where to look, and noting other contextual information about the scene etc. To model professional exploitation personnel is not an easy task. However, another approach may help at the initial approach of this challenge. One of the innovative approaches used in intelligent unmanned air system automatic piloting development can be used here too. It requires further investigation of professional exploitation personnel's behaviors, approaches, perceptions, and affections to develop better intelligent video analytics helpers, like indexing, exploitation, correlating with other context, etc. Note that with the analysis of professional UAS pilot actions at different tasks, better intelligent automatic piloting approaches emerged. In addition to providing better motion imagery analytics helpers, this further research will provide better user experience approaches and more effective human-machine interaction solutions. With this in mind, we have identified in the next section some of the key research gaps that need specific action.

4. Promising Methods

4.1 Deep Learning

The term "Deep Learning" is a new branch of machine learning enabled substantially by recent developments and advances in the field of ANNs. It aims to develop and implement algorithms that can attempt to learn multiple levels of representation of increasing complexity/abstraction to enable the machine to better reason and infer meaning from the input being processed. Although ANN is not a new concept, its usage was superseded in the 1990s by linear classifiers such as support vector machines due to their quick learning rate and higher accuracy on many supervised learning tasks, including image classification and object detection. However, due to the advances in computing power and the availability of large data sets, deep ANNs are now producing state-of-the-art results in many academic and commercial fields (Deep Learning 2015; Schmidhuber 2015). Here DL typically refers to the use of convolutional neural networks (CNNs) of several locally connected convolutional layers and fully connected classification layers implemented in high-performance computing hardware. Their strength appears to be their ability to learn hierarchical features at the different layers. When given a

large number of samples to train on, CNNs are powerful at classifying objects and events.

Hence, CNN use is of particular interest in how human learning, development, and reasoning capabilities can at some level of abstraction be conferred to machines.

Some obvious examples are automated image/vision/audio processing and NLP. These are at the core of many multimedia applications and hence critical in terms of advances in machine automation for CBA.

The conceptual purpose of applying an ANN to data is to generate a nonlinear, high order, rich representation of the data set. This is in contrast to conventional classifiers, which apply a single mapping function to convert input data into output classes. ANNs are also able to generate feature detectors and descriptors with which classification is accomplished.

When working with images, the lower layers within a convolutional network develop edge and corner detection filters that are visually similar to 2-dimensional (2-D) Gabor wavelets, a series of filters that are similar to those found in the human visual cortex. In contrast to 2-D Gabor wavelets, the first layers of a trained ANN may also provide corner, blob, and complex pattern detectors such as those found in textures. Feature complexity increases as a function of layer level. For example, a network trained on face images will develop higher order features that resemble parts of the face, such as eyes. Filters, which resemble whole faces, can be found at ever-higher levels.

ANNs provide a flexible and adaptable DL architecture that can be used for many data types. ANNs have provided accurate data classification, often doing better than competing methods in a range of areas such as biometrics, image classification, and speech recognition (Hannum et al. 2014; Lee et al. 2014; Sun et al. 2015). As an example, entries into ImageNet, an annual open image classification challenge that began in 2010, were predominantly rule-based and relied on linear classifiers. Geoff Hinton (2007) of the University of Toronto entered the first ANN-based approach used in the challenge. His entry, one of 18 entries, exhibited the lowest error by a margin of nearly 47% when compared to second place (Krizhevsky 2012). In the following year, 76% of the entries were deep-ANN-based. The ImageNet 2014 winning entry was submitted by Google, showed error rate reduction further to 6.67%, a gain of approximately 60% when compared with the first neural network deployed by Geoff Hinton's team (Szegedy et al. 2014). This provides evidence of the rapidly advancing field of DL.

As an example of application of DL to face recognition, the Chinese University of Hong Kong (CUHK) has produced state-of-the-art verification and identification

results when testing against the Labelled Faces in the Wild data set (Sun et al. 2015). The paper compares CUHK results against the world's leading commercial algorithm, NeoFace, developed by NEC Corporation of America. With respect to identification, NeoFace produced a 35% genuine acceptance rate (GAR) whereas CUHK produced an 81.4% GAR (At a 1% false acceptance rate [FAR]). In addition, the Rank-1 accuracy of NeoFace and CUHK is 56.7% and 96.0%, respectively. This is a significant improvement over leading commercial vendors. The following Table presents accuracy of DL and non-DL techniques for a range of tasks.

Table Examples of accuracy exhibited by DL and non-DL techniques for a range of tasks

Task (data set)	Non-DL	DL
Object classification (ImageNet) Krizhevsky et al. 2012)	26.6% error	6.67% error
Face recognition (local faces in the wild) (Sun et al. 2015)	35% GAR @ 1% FAR (proprietary)	81.4% GAR @ 1% FAR
Speech recognition (switchboard test set) (Goyal et al. 2013)	25.8% Word error rate	16.0% Word error rate
Optical Character Recognition (street view text data set) (McKeown et al. 1989)	0.38 F-score	0.46 F-score

A number of open source DL libraries can be used to exploit the recent research in DL: Caffe developed by Berkley Vision; Torch7 developed by multiple laboratories including Facebook, Google, Twitter, and the University of New York; and Cuda-convnet2 developed by the University of Toronto. These libraries provide relatively simple means to generate complex architectures to train on large data sets and can replicate the state-of-the-art performance found in many DL research papers. In addition the libraries leverage graphics processing unit acceleration to reduce model generation time by orders of magnitude.

Investigations into DL research, development, and applications for defense are being conducted by a number of the laboratories involved in the ET. ARL, for example, is conducting research in semantic video analytics with DL as the key element. The approach consists of building a semantic hierarchy of FMVs and sending relevant levels of information determined by network constraints, quality of information criteria, availability of computational resources, etc. Expected long-term impact is the provision of optimal semantic information to users in a timely fashion while adapting to dynamically varying computational resources (mobile devices versus mobile clouds).

It is evident from the previous section that DL and deep ANNs can offer a substantial performance improvement in terms of feature classification and lower error rates than non-DL approaches for automated image/vision/audio processing and NLP tasks. Though DL is computationally intensive, the advances in technologies and architectures in high performance computing make it a technically and commercially viable part of the solution to CBA challenges. As automated image/vision/audio processing and NLP tasks form the core of many multimedia applications, DL needs to be considered as one of the emerging areas of CBA as outlined in the following section.

4.2 Common Framework for Multimedia Analytics Development and Testing

The US Department of Defense (DOD) has invested in advanced data and decision analytics to exploit the phenomenon of big data. These efforts included WAMI, TA, and integrated information architectures. Corporate laboratories responsible for leading those efforts, respectively, are the Sensors Directorate of US Air Force Research Laboratory (AFRL), the Computational and Information Sciences Directorate of ARL, and the Massachusetts Institute of Technology Lincoln Laboratory (MIT-LL). Initially focused on developing advanced exploitation capabilities for separate text and video data, recent efforts have addressed cross-cueing and integrated processing for improved decision support. Toward that end, the AFRL-ARL-MIT-LL team led a text/video exploitation demonstration that highlighted years of advanced technology research funded by the US DOD. The event sponsor was the Naval Postgraduate School's Joint Interagency Field Experimentation (JIFX) 15–2, held 9–13 February 2015 at Alameda Island, CA.

The team's events demonstrated how the use of TA and WAMI exploitation tools could be used to support operational planning and execution. The tools exercised in the demonstration were designed to reduce cognitive workload of intelligence analysts seeking knowledge in very large document sets. Sample technologies included summarization, social network extraction, and foreign language exploitation of relations in networks, discourse and sentiment analysis, and predictions of regional conflict. Cross-cueing events between text and video were designed to trigger WAMI watch-box selections, potential individual/groups for monitoring, and contextual understanding behind observed WAMI activities. These multisource data analytics integrated research efforts are continuing with increasing emphasis on integrating previously parallel exploitation paths for text and video (Bowman and Zimmerman 2015).

4.3 Architecture

As the amount of information increases for military awareness, interoperability among national partners and sharing of fusion capabilities become paramount. NATO has been active to advance the domain of intelligence, surveillance and reconnaissance (ISR) interoperability, in particular through the Multi-Int All-source Joint ISR Interoperability Coalition (MAJIIC) program. In this context, a set of NATO standard agreements (STANAGs) have been developed and adopted to facilitate sharing of ISR information (motion and still imagery, ground moving target information, tracks, etc.) as well as intelligence products among coalition participants. The access and sharing of sensor data is made possible by the development of a distributed data storage named Coalition Shared Database (CSD) Server. For the exchange of these types of data, MAJIIC has implemented an interface defined in STANAG 4559, NATO Standard ISR Library Interface, for metadata-based access to and retrieval of archived data from any CSD throughout the interconnected MAJIIC environment (NATO OTAN 2007).

However, to make sense of a situation in complex environments (e.g., detect, recognize, understand activities, detect anomalies), there is a need to go beyond an architecture based on the definition of metadata standards for the sharing of ISR data. Research efforts are underway by various communities as part of "Big Data (BD) analytics" or "hard/soft fusion" initiatives to integrate, correlate, fuse, and analyze voluminous data from various heterogeneous sources, while considering data veracity and velocity.

Recent research in the domain of hard and soft fusion attempts to propose methods, algorithms, frameworks, and architectures to combine hard (e.g., physics-based sensors) quantitative data and soft (e.g., human-generated) qualitative data. Main categories of data sources are texts (social media, observations, intelligence reports) and sensor-based data (tracks, imagery/video). In particular, the research addresses semantic aspects such as the provision of common referencing and alignment of hard and soft data. Effective integration of multisource data improves access, discovery, correlation, better support information fusion, and ultimately provides enhanced situational awareness. In this context, Canadian research at Defence Research and Development Canada has been investigating technological solutions for the integration of multisource intelligence data in support of intelligence analysis, and has developed a prototype exploiting semantic technologies and BD. Such efforts aim at providing automated support from data ingestion to intelligence production. Data collected from heterogeneous sources, both structured and unstructured, are ingested into a unified data store according to a unified data representation scheme. The approach exploits the underlying semantics of data

sources and provides mapping (semantic alignment) using reference ontology of the domain (Boury-Brisset 2013).

For textual sources, information extraction is performed to extract significant entities related to people, organizations, events or activities these people are involved in, locations (facilities, geospatial areas), as well as relations between identified entities. Flexible indexing schemes are developed for enhanced search, and correlation between heterogeneous information. Moreover, metadata including data provenance, uncertainty, temporal, and spatial information are associated with data when available. Imagery is a very important ISR data source. In addition to metadata such as those defined in ISR-related STANAGs, image processing algorithms should be applied to support automated object/event/activity recognition, provide content-based video annotation, and thus facilitate content-based semantic search.

Consequently, data integration using a unified representation scheme, aligned with a reference ontology, facilitates information filtering, data correlation, and entity resolution.

As the data volume continues to grow, the proposed system must be scalable, able to handle real-time massive data sets, and incorporate new data resources as required. To this end, the implementation leverages BD technologies (e.g., Hadoop, MapReduce). Moreover, the platform facilitates various data mining and analytics (e.g., graph extraction, social network analysis, or visual analytics).

4.4 Common Scenarios and Data Sets

For text/imagery analysis, a scenario was constructed to focus on activities, places, groups, and people in Nigeria. The focus on Nigeria enabled an examination of a variety of different data sources and groups of people. Because of the wealth of information and reports available in the public domain, we chose the terrorist group Boko Haram as the focus for our analysis effort. As part of the scenario, we developed a thread for an improvised explosive device (IED) attack against a foreign official. Group objectives included prediction of event, location, time, sentiment analysis and discourse analysis of text, and identification of potential victim(s), and identification of specific perpetrator(s). The final goal was to provide actionable intelligence to a WAMI sensor operator.

The scenario data set that was developed for this experiment included more than 40,000 documents from 5 different sources. The “ground truth” data set included more than 700 different tactical reports, intelligence reports, police reports, speeches, tweets, and various other documents. The ground truth data set identified

relationships between groups of people and identified specific people of interest, planned meetings, and events. The ground truth was correlated to the open source data sets that were developed from reports on AllAfrica.com, *Nigeria Guardian News*, *Tribune News*, and *PM News Nigeria*. Performing groups used the data to provide insights into the planned terrorist activities, the various social groups and what their roles were, etc. In this sense, the scenario and data set provided an element of cohesion and integration across the individual technologies. This scenario and the data are unclassified.

4.4.1 Text Data Sets

Several unclassified data sets have been developed for the text analysis domain with ground truth elements. These exist in ARL at the small, medium, and large sizes and include the Ali Baba, Kandahar, and AllAfrica.com data sets. The Ali Baba data set contains approximately 600 messages that combine to detect an IED event (Mittrick et al. 2012). The Kandahar data set was constructed around a social network extraction and analysis set of challenges and includes approximately 800 messages. This is a much richer data set and includes 9 subsets of networks, main topics of discussion in the various networks, sentiment toward topics, and topic-person pairing (e.g., which topics are of primary concern to an individual/group?). The AllAfrica.com data set includes more than 40,000 documents. Each of these data sets contains ground truth.

The development of data sets is dominated by US research, through various government agencies, due to higher levels of investment and greater volume of projects compared with other NATO allies. In some cases, this has been through collaboration with academia and industry, including those from other nations, and has resulted in some data sets that can be shared. Outline of some of these developments is described in the following section, including the opportunities for sharing the data sets. Other nations involved in the ET also have data sets and supporting frameworks for experimentation and testing and should be considered in further research but are not detailed here.

4.4.2 Video Data Sets

Video data sets dedicated to human action and activity recognition have been created to allow researchers to compare different recognition systems with the same input data. “Computer Vision and Image Understanding” provides a survey of public data sets (68 data sets are reported in the authors), and some guidance of the most suitable data set for benchmarking their algorithms (Chaquet et al. 2013).

These data sets often define categories of actions/events to be exploited by recognition algorithms. Among them, the DARPA VIRAT data set has been

produced to provide realistic and challenging data for video surveillance in terms of resolution (wide range of resolution and frame rates), background clutter, diversity in scenes, camera viewpoints (ground and aerial videos), and human activity/event categories. Diverse types of human actions and human/vehicle interactions are included (23 event types).

The video analytics algorithms suffer from a lack of context. For example, for counter-IED, it is relevant to know when there are activities close to a road that may point to somebody placing the device. Yet, most of the activities around the road are normal, such as travelers who take a rest and people selling goods. It is infeasible to endow an algorithm with a complete model of which activities may happen around that particular road at each date, time, weather, and for each type of actor. This makes the algorithm incomplete, which will lead to errors. Textual metadata that accompanies the video data may guide the algorithm to improve its assessment and consequently reduce its errors. For instance, prior information about a suspect person like clothing and vehicle will increase the performance because it provides focus in the midst of many other people. Improving video analytics algorithms by incorporating external, textual metadata is a novel research area where significant impact on military usability is expected.

At the very initial step where motion imagery is captured (mostly by UASs), intelligent capturing and initial processing sensory systems could further be investigated. Since in military domain each planned intelligence task has some objectives, stages, and related metadata to be assigned at various parts of the produced video, initial video indexing and key frame information could also be produced as the task is taking place with additive UAS intelligence personnel's audio and metadata entries. Sensory motion behaviors (zoom to a location for a period of time, change of focus to another location, or moving between different interest points etc.) could also be recorded for further hints for the indexing of the important activities in the motion imagery.

Approaches to index motion imagery depend on the predefined domain differentiators, like face appearance, motion detection, and event occurrence (such as a truck parking at some point, etc.) Indexing primarily helps exploitation of motion imagery, providing some important instants in a long sequence. Indexing helps fast retrieval of motion imagery according to the metadata provided by hierarchical indexing information. However, further research is still required for hierarchical indexing where indexing is obtained by using domain-specific hierarchical semantic identifiers. Further investigation and scenarios need to be developed for producing effective motion imagery exploitation. Human-machine interaction research in the field of hierarchically indexed motion imagery should be pursued, where domain-specific hierarchical semantic identifiers and hierarchical

indexing of motion imagery is effectively presented to the intelligence personnel. Further research is also needed on motion-based index generation, where different types of motion are taken into account (background motion of static structures related with UAS flight, background motion generated by normal patterns such as traffic flow, an explosion, and its aftereffects at a location, etc.).

5. Conclusions/Recommendations

In the area of DL, there is a significant amount of research in both the academic and commercial sectors into large-scale machine learning, which has produced state-of-the art results as outlined in previous sections. The capabilities are now commonly used for commercial use for image processing (Google, Baidu), audio processing (Google, Apple, and Microsoft) and the financial sector. There are many defense and security challenges that could benefit from these techniques, particularly when triaging large data sets (e.g., image exploitation, biometrics and security, speech analysis) (Howe 2015).

As an emerging research area for CBA, and from a defense and security perspective, it is recommended that DL research is pursued further on 2 fronts, applications and system architectures.

As we consider applications, further research and performance measurements are needed, especially false alarm rates with large data sets particular to defense and security. This is using open source toolkits, thereby reducing risk through using established approaches, but applying them to process text, images, and audio to enable better

- Face recognition;
- Object detection in imagery/video;
- Speech recognition; and
- Optical character recognition.

The integration of these areas to handle multimedia streams in parallel and the fusion of the output layers would be a worthwhile and stretching research challenge.

As we consider system architectures, further research is needed to advance DL techniques through improving the mathematical foundations of current algorithms as well as developing new approaches for their implementation and new algorithms that can better exploit advances in high-performance computing. System co-design is needed, algorithms and hardware, to be able to engineer and integrate system

architectures that are optimized for CBA and DL in general. Technical areas include, but are not limited to, the following:

- Hardware for DL, including neuromorphic chips and low-power, high-performance field programmable gate arrays
- New optimization methods across the convolutional layers and back propagation techniques
- Training with fewer examples
- Distributed learning across systems

An overall research goal would be how to integrate with and best exploit the current and future commodity distributed systems (e.g., servers, cloud) and distributed services (e.g., computation, storage) expected to be available to an enterprise. Included in this goal is how those services would be extended to support collaborative endeavors and leverage the distributed systems capability for research, development, and experimentation.

The work of the exploratory team has shown, within the context outlined in the previous section, that advances in CBA research have the potential to be key enablers for defense and security analysts and military experts to exploit data from multiple sources in a rapid fashion for sensemaking, decision support, and knowledge generation. This is especially true in the complex, dynamic and changing defense and security environments presented to NATO allies, such as hybrid warfare.

A huge improvement for video analytics is expected when contextual information can be successfully incorporated such that the algorithm knows when and where to focus on a subset of the scene and actors. This would result in more accurate predictions because a large part can be ignored by the algorithm and it can focus on a few candidate persons and/or parts of the scene. The most promising improvement seems to incorporate textual metadata, which often accompanies military video data. Section 3, Research Gaps, provides explanation and motivation to further pursue this research direction.

It is recommended that further NATO collaborative research be undertaken with specific focus on maturing specific research elements and specific research gaps identified by the exploratory team and outlined in this report. This, in summary, is enhanced real-time analytics of heterogeneous multimedia streams (image, video, text, speech etc.) enabled by enhancements in the contextual understanding of complex events through advances in computational/human processing techniques. It is also event prediction by the automated extraction of spatio-temporal features,

hidden clusters, network structures, and resource flows. Finally, it will aid in the application of better machine-learning/DL algorithms and processing for the automated translation, parsing, information extraction, and summarization of unstructured and semistructured data from multiple streams.

It is suggested that the following technical aspects should be the focus of the research:

- Capture and index motion imagery; further investigate intelligent capturing and initial processing by sensor systems, to include initial video indexing and key frame information produced in audio and metadata entries.
- Exploit imagery indexing through hierarchical methods using semantic identifiers and human evaluations of exploitation results.
- Explore motion-based index generation to generate rapid and robust retrieval of context. Types of motion include background motion of static structures related with sensor flight, background motion generated by normal patterns such as traffic flow, an explosion and after effects at a location, etc.
- Expand the DL approach for semantic video analytics through a semantic hierarchy of FMV. Long-term impact is the provision of optimal semantic information to users in rapid fashion while adapting to dynamically varying computational resources.
- Explore the mechanisms by which text analysis results can be used to drive/exploit video and imagery indexing and retrieval.
- Explore frameworks for optimizing multimedia analytics via systems engineering and architectural design concepts.

These types of complex analyses and advancement of approaches cannot be done in isolation and would benefit significantly from the technical and financial gearing afforded by collaboration between NATO allies. Furthermore, the greatest value and utility of the outputs and outcomes of the research, as well as their critique, will be achieved through such collaboration.

NATO and coalition military leaders, commanders, and intelligence analysts need interoperable tools that cross-cue knowledge obtained from one method to generate taskings in another. The recommended research focus should help build the cross-cued solution from advances in content-based multimedia data analytics. If the research is successful, it will significantly improve NATO abilities to generate knowledge from extremely large stores of text, imagery, and video caches to speed situational awareness and decision making.

6. References

Antol S, Aishwarya A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. VQA: Visual question answering. Ithaca (NY): Cornell University; 2015 [accessed 2016 June 20]. http://arxiv.org/abs/1505.00468. arXiv:1505.00468.

Bird S, Klein E, Loper E. Natural language processing with python. Sebastopol (CA): O'Reilly Media, Inc.; 2009.

Blaschko M, Girshick R, Kannala J, Kokkinos I, Mahendran S, Maji S, Mohamed S, Rahtu E, Saphra N, Simonyan K, Taskar B, Vedaldi A, Weiss D. Towards a detailed understanding of objects and scenes in natural images. Baltimore (MD): Johns Hopkins University, Center for Language and Speech Processing; 2012 Aug 8.

Boury-Brisset A-C. Managing semantic big data for intelligence. In: Proceedings of the International Conference Semantic Technologies for the Intelligence, Defence and Security (STIDS); 2013 Nov 12–15; Fairfax, VA.

Bowman EK, Zimmerman RJ. US Army Research Laboratory joint interagency field experimentation 15-2 final report. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2015 Dec. Report No.: ARL-TR-7562.

Chaquet JM, Carmona EJ, Fernández-Caballero A. A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding. 2013;117:633–659.

Crevier D, Lepage R. Knowledge-based image understanding systems: a survey. Computer Vision and Image Understanding. 1997;67.2:161–185.

Deep learning. Wikipedia, The Free Encyclopedia. [accessed 2015 Sep]. https://en.wikipedia.org/wiki/Deep_learning.

Duda R, Hart P, Stork D. Pattern classification. 2nd ed. New York (NY): Wiley-Interscience; 2000.

Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. New York (NY): Cambridge University Press; 2006.

Fisher K, Counts S, Kittur A. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI ′12; 2012 May 5–10; Austin, TX. New York (NY): ACM; c2012. p. 247–256.

Garcia-Gasulla D, Béjar J, Cortés U, Ayguadé E, Labarta J. Extracting visual patterns from deep learning representations. Ithaca (NY): Cornell University; 2015 [accessed 2016 June 20]. http://arxiv.org/abs/1507.08818. arXiv:1507.08818.

Girju R. Automatic detection of casual relations for question answering. In: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering. MultiSumQA ′03; 2003 Jul 11. Stroudsburg (PA): Association for Computational Linguistics; c2003. p. 76–83.

Goyal N, Leshed G, Fussell SR. Leveraging partner's insights for distributed collaborative sensemaking. CSCW 2013. Proceedings of the 2013 ACM conference on computer supported cooperative work companion; 2013 Feb 23–27; San Antonio, TX. New York (NY): ACM; c2013. p. 15–18.

Grant K. Dimensions of collaborative work. Presented at: Human Computer Interaction Consortium. 2001 Feb 7–11; Fraser, CO [accessed 2016 June 20]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.9832&rep=rep1&type=pdf.

Grimmer J, Stewart B. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Political Analysis. 2013;21:267–297. doi: 10.1093/pan/mps028.

Hannum A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, Ng AY. DeepSpeech: scaling up end-to-end speech recognition. Ithaca (NY): Cornell University; 2014 [accessed 2016 June 20]. http://arxiv.org/abs/1412.5567. arXiv:1412.5567.

Heer J, Argrawala M. Design considerations for collaborative visual analytics. Information Visualization. 2008;7:49–62.

Hinton G. Learning multiple layers of representation. Trends in Cognitive Sciences. 2007;11(10):428–434.

Howe J. Benefits of deep learning for military and security requirements Porton Down (UK): Defence Science and Technology Laboratory (DSTL); 2015. Report No.: DSTL TR-86897.

Keel PE. EWall: A visual analytics environment for collaborative sensemaking. Information Visualization. 2007;6:48–63.

Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. NIPS 2012; 2012 Dec 3–8; Lake Tahoe, NV.

Proceedings of the Neural Information Processing Systems 2012; La Jolla (CA): NIPS. c2012. p. 1097–1105.

Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. Ithaca (NY): Cornell University; 2014 [accessed 2016 June 20]. http://arxiv.org/abs/1409.5185. arXiv:1409.5185.

Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P. Microsoft COCO: common objects in context. Ithaca (NY): Cornell University; 2015 [accessed 2016 June 20]. http://arxiv.org/abs /1405.0312. arXiv:1405.0312.

Mao J, Xu W, Yang Y, Wang J, Yuille A. Deep captioning with multimodal recurrent neural networks (m-RNN). Ithaca (NY): Cornell University; 2015 [accessed 2016 June 20]. https://arxiv.org/abs/1412.6632v1. arXiv:1412.6632.

McKeown DM Jr, Harvey WA, Wilson A, Wixson LE. Automating knowledge acquisition for aerial image interpretation. Computer Vision, Graphics, and Image Processing. 1989;46.1:37–81.

Mittrick M, Roy H, Kase S, Bowman E. Refinement of the Ali Baba data set. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2012. Report No.: ARL-TN-0476.

North Atlantic Treaty Organization (NATO) Organisation du Traité de l'Atlantique Nord (OTAN). MAJIIC factsheet. Brussels (Belgium): North Atlantic Treaty Organization.; nd. http://www.nato.int/docu/update/2007/pdf/majic.pdf.

Paul SA, Morris MR. Sensemaking in collaborative web search. In: Daniel Russell D, Pirolli P, editors. Human Computer Interaction Special Issue on Sensemaking. 2011;26(1):38–71.

Pratikakis I, Bolovinou A, Gatos B, Perantonis S. Semantics extraction from images. In: Paliourus G, Spyropoulos CD, Tsatsaronis, G, editors. Knowledge-driven multimedia information extraction and ontology evolution. Heidelberg (Germany): Springer Berlin Heidelberg; 2011. pp. 50–88.

Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks. 2015;61:85–117. doi: 10.1016/j.neunet.2014.09.003.

Shah A, Baum S, Dwivedi V. Neural substrates of linguistic prosody: evidence from syntactic disambiguation in the productions of brain-damaged patients. Brain and Language. 2006(96):78–89. doi:10.1016/j.bandl.2005.04.005.

Sun Y, Liang D, Wang X, Tang X. DeepID3: face recognition with very deep neural networks. Ithaca (NY): Cornell University; 2015 [accessed 2016 June 20]. http://arxiv.org/abs/1502.00873. arXiv:1502:00873.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. Ithaca (NY): Cornell University; 2014 [accessed 2016 June 20]. http://arxiv.org/abs/1409.4842. arXiv:1409:4842.

Tate A, Buckingham Shum S, Dalton, J, Mancini C, Selvin AM. Co-OPR: design and evaluation of collaborative sensemaking and planning tools for personnel recovery. Edinburgh (UK): University of Edinburgh; 2006. Report No.: KMI-TR-06-07.

Umpathy K. Requirements to support collaborative sensemaking. Presented at: CSCW CIS Workshop ′10; 2010 Feb 7; Savannah, GA.

Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. Ithaca (NY): Cornell University; 2015 [accessed 2016 June 20]. http://arxiv.org/abs/1411.4555. arXiv:1411.4555.

Wallace J, Scott S, MacGregor CG. Collaborative sensemaking on a digital tabletop and personal tablets: prioritization, comparisons, and tableax. CHI 2013. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2013 Apr 27–May 3; Paris, France. New York (NY): ACM. c2013. p. 3345–3354.

Xu R, Xiong C, Chen W, and Corso J. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. Proceedings of Twenty-ninth AAAI Conference on Artificial Intelligence; 2015 Jan 25–30; Austin, TX. Palo Alto (CA): AAAI Press. c2015. p. 2346–2352.

Yao BZ, Yang X, Lin L, Lee MW, and Zhu S-C. I2t: Image parsing to text description. Proceedings of the IEEE 2010;98.8:1485–1508 [accessed 2016 June 20]. http://www.stat.ucla.edu/~sczhu/papers/I2T_IEEE_proc.pdf.

List of Symbols, Abbreviations, and Acronyms

2-D	2-dimensional
AFRL	US Air Force Research Laboratory
AI	artificial intelligence
ANN	Artificial Neural Network
ARL	US Army Research Laboratory
BD	Big Data
CBA	Content-Based Analytics
CNN	convolutional neural networks
CS	collaborative sensemaking
CSD	Coalition Shared Database
CUHK	Chinese University of Hong Kong
DARPA	Defense Advanced Research Projects Agency
DEFT	Deep Exploration and Filtering of Text
DL	deep learning
DOD	Department of Defense
DSTL	Defence Science and Technology Laboratory
ET	exploratory team
FAR	false acceptance rate
FMV	full-motion video
GAR	genuine acceptance rate
HARVEST	Human Activity Recognition in VidEo STreams
IARPA	US Intelligence Advanced Research Projects Activity
IE	information extraction
ISR	intelligence, surveillance, and reconnaissance
IST	Information Systems Technology

JIFX	Joint Interagency Field Experimentation
MADCAT	Multilingual Automatic Document Classification, Analysis and Translation
MIT-LL	Massachusetts Institute of Technology Lincoln Laboratory
MOD	Ministry of Defence
NATO	North Atlantic Treaty Organization
NLP	natural language processing
OSINT	Open Source Intelligence
OWL	web ontology language
POL	pattern of life
RTG	Research Technology Group
SAS	System Analysis and Studies
STANAG	standard agreement
SUBTLE MURI	Situation Understanding Bot through Language and Environment Multi University Research Initiative
TA	text analytics
TNO	Netherlands Organisation for Applied Scientific Research
UAS	unmanned autonomous system
VIRAT	Video Image Retrieval and Analysis Tool
VMR	Visual Media Reasoning
WAMI	wide area motion imagery

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

2 (PDF) DIRECTOR US ARMY RESEARCH LAB RDRL CIO L IMAL HRA MAIL & RECORDS MGMT

1 (PDF) GOVT PRINTG OFC A MALHOTRA

1 (PDF) DIRECTOR US ARMY RESEARCH LAB RDRL CII T E BOWMAN

INTENTIONALLY LEFT BLANK.